

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Andreas SAVVA

Application No.:

Group Art Unit: Not Yet Assigned

Filed: November 26, 2003

Examiner: Not Yet Assigned

For: THE APPARATUS AND THE METHOD FOR INTEGRATING NICS WITH RDMA
CAPABILITY BUT NO HARDWARE MEMORY PROTECTION IN A SYSTEM WITOUT
DEDICATED MONITORING PROCESSES

**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN
APPLICATION IN ACCORDANCE
WITH THE REQUIREMENTS OF 37 C.F.R. § 1.55**

Commissioner for Patents
PO Box 1450
Alexandria, VA 22313-1450

Sir:

In accordance with the provisions of 37 C.F.R. § 1.55, the applicant(s) submit(s) herewith
a certified copy of the following foreign application:

Japanese Patent Application No(s). 2002-357449

Filed: December 10, 2002

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing
date(s) as evidenced by the certified papers attached hereto, in accordance with the
requirements of 35 U.S.C. § 119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: November 26, 2003

By: 

J. Randall Beckers
Registration No. 30,358

1201 New York Ave, N.W., Suite 700
Washington, D.C. 20005
Telephone: (202) 434-1500
Facsimile: (202) 434-1501

JAPAN PATENT OFFICE

This is to certify that the annexed is a true copy of the following application as filed with this office.

Date of Application: December 10, 2002

Application Number: Patent Application No. 2002-357449
[ST.10/C] [JP2002-357449]

Applicant(s): FUJITSU LIMITED

September 25, 2003

Commissioner,

Japan Patent Office Yasuo IMAI

Certificate No. P2003-3078526

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 2 年 1 2 月 1 0 日
Date of Application:

出 願 番 号 特 願 2 0 0 2 - 3 5 7 4 4 9
Application Number:
[ST. 10/C]: [J P 2 0 0 2 - 3 5 7 4 4 9]

出 願 人 富 士 通 株 式 会 社
Applicant(s):

2 0 0 3 年 9 月 2 5 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康



【書類名】 特許願

【整理番号】 0252079

【提出日】 平成14年12月10日

【あて先】 特許庁長官殿

【国際特許分類】 H04L 29/08

【発明の名称】 R D M A 機能を持った N I C をハードウェアメモリ保護
を行わないで、専用のモニタプロセスなしにシステムに
組み込むための装置

【請求項の数】 5

【発明者】

【住所又は居所】 神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号 富士通
株式会社内

【氏名】 アンドレアス サバ

【特許出願人】

【識別番号】 000005223

【氏名又は名称】 富士通株式会社

【代理人】

【識別番号】 100074099

【住所又は居所】 東京都千代田区二番町 8 番地 2 0 二番町ビル 3 F

【弁理士】

【氏名又は名称】 大菅 義之

【電話番号】 03-3238-0031

【選任した代理人】

【識別番号】 100067987

【住所又は居所】 神奈川県横浜市鶴見区北寺尾 7 - 2 5 - 2 8 - 5 0 3

【弁理士】

【氏名又は名称】 久木元 彰

【電話番号】 045-573-3683

【手数料の表示】

【予納台帳番号】 012542

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9705047

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 R D M A 機能を持った N I C をハードウェアメモリ保護を行わないで、専用のモニタプロセスなしにシステムに組み込むための装置

【特許請求の範囲】

【請求項 1】 R D M A 機能を有した、ホストと複数のホストからなるネットワーク内の該ホストに設けられた装置であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第 1 のメッセージを、該ネットワーク内の全ての該複数のホストに送信する手段と、

該複数のホストからの該ホストへの R D M A アクセスを利用不可能とする手段と、

該ホストに第 2 のメッセージを送信することによって、第 1 のメッセージに回答する手段と、

該ホストが、該複数のホストからの R D M A アクセスを受付可能であることを示す第 3 のメッセージを、該複数のホストの全てから第 2 のメッセージを受信し、R D M A 機能を利用可能とした後、該複数のホストの全てに送信する手段と、を備えることを特徴とする装置。

【請求項 2】 前記装置は、前記ホストのドライバに含まれることを特徴とする請求項 1 に記載の装置。

【請求項 3】 他のホストに R D M A アクセスを行うための情報を有する翻訳保護テーブル手段を更に有し、

前記第 1 のメッセージを受信したとき、該第 1 のメッセージを送信したホストに関する情報を該翻訳保護テーブル手段からクリアし、該ホストへの R D M A アクセスを不可能にすることを特徴とする請求項 1 に記載の装置。

【請求項 4】 前記第 2 のメッセージは、アクノレジメント、ノンアクノレジメント、及び、前記複数のホストから送られた前記第 1 のメッセージの一つであり、ノンアクノレジメントは、ハードウェアによって生成されることを特徴とする請求項 1 に記載の装置。

【請求項 5】 前記ホストは、R D M A 機能と他のメッセージ通信機能を有す

るネットワークインターフェースカードを有し、R D M A 機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする請求項 1 に記載の装置。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、R D M A 機能を有するホストマシンをネットワークに組み込み、あるいは、再組み込みするための装置に関する。

【 0 0 0 2 】

【従来の技術】

リモートDMA (R D M A : Remote Direct Memory Access) は、ネットワークインタフェースカード (N I C : Network Interface Card) に対する、特に、R D M A 拡張機能を通常のA P I に付加するという最近の提案においては、重要な要求となってきた。R D M A をサポートするN I C は、ローカルメモリへのR D M A アクセスが適切に有効化することを確保するために、メモリプロテクション、例えば、翻訳保護テーブル (T P T : translation and protection table) をサポートすることが要求される。このようなテーブルを実装することは、ハードウェアコストを増大し、特に、周辺バス (例えば、P C I : Peripheral Component Interconnect) へ接続されるN I C の性能に影響を与える可能性がある。ハードウェアベースのT P T の代替法法としては、他のN I C へのR D M A アクセスを有効化するための機能を果たすN I C ドライバを構成することである。この場合、全体のシステムは、R D M A アクセスがホストをクラッシュさせるようには生じないことを確保するために協同しなくてはならない。

【 0 0 0 3 】

この問題に対する解決方法は、ホスト上で動作するドライバにコピーが保持される、システム全体をカバーする保護テーブルの形で実現されるであろう。ドライバは、テーブルのローカル部を更新し、変更を遠隔の度ラバに反映する。ホストが再起動しない限りは、問題はない。しかし、ホスト、例えば、H O S T _ A が再起動された場合には、システム内の残りのホストは、すぐには、このことを認知しない、あるいは、認知できない。H O S T _ A が、R D M A アクセスを再

起動の直後に可能とすると、保護テーブル内の古い情報（H O S T _ A が再起動される前）に基づいたリモートホストから始められた R D M A 動作は、H O S T _ A をクラッシュさせることがある。

【 0 0 0 4 】

例えば、ホスト（ネットワーク内のノード）が再起動されたとき、ホストの内部 R D M A 設定、すなわち、登録されたメモリ領域は解除され、初期化される。しかし、他のホストの R D M A 設定が更新されていないと、他のホストは、R D M A を使って、最近再起動したホストのメモリにアクセス使用とするかもしれない。このとき、当該ホストのメモリ割当ては、再起動後初期化されており、再起動されたホストのメモリ割当ての初期化が他のホストの設定に反映されていないので、他のホストの再起動したホストのメモリへの、古い設定によって決定されたアドレスへのアクセスは、再起動したホストの動作に重要な情報を上書きしてしまうかもしれず、再起動したホストをクラッシュさせるかもしれない。

【 0 0 0 5 】

従って、再起動されたホストが他のホストによってそのように認識され、再起動したホストに関連するシステムの保護テーブル内の情報が、再起動したホストが完全にシステムに再び組み込まれるまで、無効化されることを確保することが本質的である。

【 0 0 0 6 】

【発明が解決しようとする課題】

通常の解決方法は、おそらくハートビートユーザレベルプロセスと共に、リモートホストの状態の変化をドライバに通知する各ノード上のユーザレベルプロセス（U n i x（登録商標）におけるデーモンプロセスのような）の形を採用するかもしれない。例えば、システムは、各ホストに N I C 制御プロセスとハートビートプロセスを有する。N I C 制御プロセスは、N I C を初期化する機能を持ち、N I C ドライバに、システムにおける例えば、ホストの再起動などの変化を通知する。各ハートビートプロセスは、ハートビートプロセスとメッセージを交換することによって、他のホストの状態を監視し続ける。例えば、ホスト、H O S T

__A が、落ちた場合には、ハートビートメッセージは送られない。所定の時間後、他のハートビートプロセスは、HOST__A が落ちたことを断定する。各ハートビートプロセスは、NIC 制御プロセスにそのことを通知し、これは次にドライバに通知して、HOST__A への全てのアクセスをブロックする。HOST__A が復帰すると、そのハートビートプロセスは再びメッセージを送信し始める。他のハートビートプロセスは、メッセージを受け取り、これらの NIC 制御プロセスに、これを通知し、こんどは、ドライバにこれを通知して、HOST__A へのアクセスを再び可能とする。しかし、このアプローチには幾つかの問題がある。例えば、NIC の初期化は、ユーザレベルプロセスである NIC 制御プロセスに依存しているため、NIC が初期化されるまでに時間がかかる。ホストの負荷が大きい場合には、このプロセスは、ちょうど良く動作するようにはスケジューリングされいなくてもいい（システム応答が遅い）。この間、NIC は使用できない。また、NIC 制御プロセスあるいはハートビートプロセスのいずれかが落ちるかもしれない。NIC 制御プロセスが落ちることは、システムイベントに対するホストの応答を含み、例えば、再起動したホストへのアクセスが長時間再開されないかもしれない。ハートビートプロセスが落ちることは、ホストのクラッシュと間違えられる可能性があり、従って、全体のシステム性能に影響する。すなわち、ユーザレベルプロセスを使用することは、システム全体の信頼性を低くすることになる。

【0 0 0 7】

他の問題は、潜在的に、RDMA リクエストによって使用されるものとは異なる、組み込みプロトコルのための通信経路を用い、NIC の RDMA 機能が利用可能とされる前に、実行途中の RDMA アクセスが無いことを確保することができないユーザレベルプロセスに起因することになる。

【0 0 0 8】

ノードが再起動されたとする。そのノードのドライバは、以前の状態の全ての情報を失う。特に、全ての、以前に登録したメモリ領域と、他のノードに登録されているメモリ領域に関する全ての情報を失う。しかし、リモートノードは、そのノードが再起動されたことを知らないかもしれない。あるドライバが当該ホス

トが再起動されたことを知り、他のドライバが知らないようなシステムに再起動されたホスト上のNICを再組み込みする問題は、解決する必要がある。ターゲットであるNICに対して、動作途中であるRDMA動作が無いことを確保しなければならないという意味において、解決方法は安全でなくてはならない。もし、そのようなRDMA動作が有ると、これらは、ホストが再起動される前に始動されており、古い情報に基づいており、ホストをクラッシュさせる可能性がある。再組み込みの終わりには、そのNICドライバは、システムの保護テーブルの最新のコピーを有していなければならない、システムの一部である、他の全てのホストは、そのNICがシステムの一部として機能していることを知らなければならない。この問題をユーザレベルモニタプロセスで解決することは、安全な時間になるまで、NICの全体の初期化を送らせることになる。ユーザレベルプロセスが通信するためには、各ホストは、他のNICと、RDMA機能なしに、適合されていなければならない、全体のハードウェアコストを増加させる。

【0009】

本発明の課題は、RDMA機能を有するNICを持つホストを、NICの初期化を制御するのに、ユーザレベルプロセスに頼らないで、ネットワークに安全に組み込むための装置を提供することである。

【0010】

【課題を解決するための手段】

本発明の装置は、RDMA機能を有している複数のホストからなるネットワーク内のホストにある装置であって、ネットワーク内のホストが起動されたとき、ネットワーク内の複数のホストの全てに、当該ホストが起動されたことを示す第1のメッセージを送信する手段と、複数のホストから当該ホストへのRDMAアクセスを全て利用不可能にする手段と、当該ホストに第2のメッセージを送ることによって、第1のメッセージに応答する手段と、当該ホストが複数のホストからのRDMAアクセスを受け付ける準備ができたことを示す第3のメッセージを、複数のホストの全てから第2のメッセージを受け取り、RDMA機能が利用可能とされた後に、全ての複数のホストに送信する手段とを備えることを特徴とする。

【0011】

本発明によれば、ネットワーク内のホストが起動される毎に、ホストは、ネットワークの全てのホストにこれを通知する。従って、全てのホストは、どのホストが起動されたかを知ることができ、当該ホストへのRDMAアクセスに使用される情報を、正しくクリアし、更新することができる。結果として、起動されたホストは、ネットワークに安全に組み込み、あるいは、再組み込みされるが、クラッシュは生じない。起動された、あるいは、再起動されたホストのネットワークへの組み込み、あるいは、再組み込みは、同様に行うことができる。

【0012】

このアプローチの独自性の一部は、SEND/RECEIVE機能と、RDMA機能の初期化を分離したことにある。SEND/RECEIVE初期化は、NIC自身がそのためのプロトコルを使って使用されるように実行され、従って、全体のハードウェアコストを下げる。RDMA機能を有しない他のNICが当該プロトコルを実行する必要はない。更に、NIC自体がこのプロトコルを実行するのに使われるため、プロトコルが終了すれば、システム内にRDMAアクセスが処理されずに残っていることがないことの保証を、ユーザプロセスが使用される場合よりも簡単に達成することができる。更に、このプロトコルが依然処理中であっても、SEND/RECEIVEを介して、通常の通信をNICを持ち手開始することが可能である。また、ユーザプロセスに対する依存性を減少させることで、システム内の部品数を減らすことができ、全体の信頼度を上げることができる。この機能をカーネルに含ませることにより、ドライバの開発コストは上がるが、他のコストは、プロトコルが簡単であるため、かなり小さくなる。

【0013】

【発明の実施の形態】

ここで、再起動されたホストのNICをシステムに安全に再組み込みし、これらのRDMA機能を利用可能にすることを確保するためにユーザレベルプロセスを使用しないアプローチを示す。このアプローチは、システム内のNICドライバ間の協同にのみ依存している。この機能をドライバに実装することは、開発コストを増加させるが、高速な応答を可能にするという利点がある。（ドライバは

、他のノードからの入力プロトコルメッセージに応答するのに、割り込みの処理を利用することができる。)更に、ドライバは、そのRDMA機能を利用可能とする前に、処理途中である古いRDMAアクセスが無いことを確保することができる。要求されるロジックは、ユーザレベルプロセスによって要求されるものより、おそらくは簡単で、従って、信頼性も改善されるであろう。

ーシステム構成

図1は、本発明の実施形態が基礎とするシステム構成を示す図である。

【0014】

システムは、多くのホスト10-1~10-7によって構成されている。各ホスト10-1~10-7は、1以上のNIC11を持っている。NIC11は、ネットワーク12を介して接続される。NIC11は、別々に、SEND/RECEIVE機能11-1とRDMA機能11-2の両方を持つ。例えば、SEND/RECEIVE機能11-1を利用可能とし、同時にRDMA機能11-2を利用不可能とすることが可能である。

【0015】

NIC11は、専用のハードウェアのアドレス翻訳保護テーブル(TPT)を持たない。そのかわり、ドライバ13は、協同して、ソフトウェアのシステム全体に渡る保護テーブル13-1を持つ。テーブル13-1は、ローカル部とリモート部の2つの部分からなる。ローカルなユーザアプリケーションがドライバ13にメモリを登録する結果として、テーブルのローカル部にエントリが加えられる。ユーザプロセスは、登録したメモリへのRDMAアクセスを利用可能としたり利用不可能とすることができる。ドライバ13は、テーブル13-1のローカル部の変更を他のノードのテーブル13-1の対応するリモート部に反映させる。再起動するホストが無い限り、全てのドライバ13は、保護テーブル13-1に同じ情報を持っている。

【0016】

各ドライバの保護テーブル13-1は、システム内の全てのNIC11についての情報を含んでいる。ローカルプロセスがRDMA動作を発行すると、ドライバ13は、保護テーブル13-1のコピーをチェックし、リモートアクセスが許

可されているか否かを判断する。(保護テーブルを他の形で実現することも可能である。例えば、ドライバは、各RDMAトランザクションを異なるプロトコルで有効化する。この場合、各ドライバは、ローカルアドレスのみを登録するテーブルを持ち、リモートアクセスが要求されたときには、リモートアクセスの要求が発行される前に、プロトコルを確立するメッセージが、ローカルアドレスのみを有するテーブルを使って、ホスト間でやりとりされる。しかし、古い保護情報を使用するという問題は依然存在する。以後に示すアプローチは依然適用可能である。)

システム内のNIC 11の最大数は、全てのホスト10-1~10-7に知られているか、確定可能である。この最大数は、現在接続されているNIC 11データはなく、ネットワーク12に接続可能なNIC 11の可能な全数を反映している。メッセージあるいは、RDMA動作は、NIC 11が適切に初期化されていないあるいは、ネットワーク12に接続されていないのでない限りは、データを配布することになる。後者の場合、メッセージを送信したドライバ13が見ることができる、ハードウェアで生成された、ネガティブアクノレジメント(NACK)がある。

【0017】

ドライバ11がNACKを検出すると、NIC 11がシステムに再組み込みされるまで、リモートNIC 11へのアクセスをドライバが禁止する。

最後に、同じ送信元、宛先のRDMA及びSEND動作は、同じ経路を使用する。ネットワーク12は、このような動作が互いに追い越すことを許さない。

【0018】

図2は、ホストの構成を示す図である。

ホスト10-iと10-jは、NIC 11を介して、ネットワーク12に接続される。ホスト10-iと10-jは、ユーザ空間とカーネル空間を有している。プロセスPは、ユーザ空間に有る。ドライバは、カーネル空間にある。NICは、ホスト10-iと10-jのハードウェアの一部である。各NIC 11は、ネットワーク12とドライバ13のインターフェースを持ったSEND/RECEIVE機能11-1とRDMA機能11-2を有している。

【0019】

プロセスPが起動されると、プロセスPは、ドライバ13のRDMA保護機能内の翻訳保護テーブル(TPT)13-1にプロセスP自身が使用するメモリ領域iを登録する。そして、TPT13-1は、プロセスPに割り当てられた領域iのエントリを有する。同様に、ホスト10-jにおいては、プロセスMは、メモリ領域Xをドライバ(不図示)に登録する。この登録は、ホスト10-iのTPT13-1に反映される。そして、ホスト10-iのTPTは、プロセスMの領域Xのエントリを持つようになる。

【0020】

プロセスPがプロセスMにアクセスする必要があるときは、プロセスPは、TPT13-1の領域Xのエントリを参照して、アクセスのための要求を発行し、アクセスが確立される。プロセスMがクラッシュしたか、ホスト10-jがクラッシュした場合には、TPT13-1の領域Xのエントリは、ホスト10-jのプロセスMが利用できないことを示すために、TPT13-1から消去される。

【0021】

図3は、翻訳保護テーブルの構成を示す図である。

図3においては、ネットワーク内に、0～N-1まで番号付けられたN個のノードがあるとする。

【0022】

TPTは、上記したように、ローカル部とリモート部を有している。ノード0では、TPTのローカル部に、自ノード0のためのメモリ領域が登録されている。リモート部には、ノード1～ノードN-1のためのメモリ領域が登録されている。TPTのエントリフォーマットが図3(b)に示されている。TPTの各エントリは、項目として、論理アドレス、物理アドレス、長さ、保護属性を有している。ユーザ空間のユーザプロセスは、論理アドレスを用いて、アクセス要求を発行する。この論理アドレスは、TPTを参照することによって、物理アドレスに変換され、物理アドレスを用いて、宛先にメッセージを送信する。長さの項目は、物理アドレスの項目で特定されるアドレスの領域の長さを、例えば、バイト単位で示すものである。保護属性の項目は、物理アドレスで示される領域の、メ

メモリ保護の種類（読み取り専用、書き込み専用など）を特定するものである。

【 0 0 2 3 】

図 4 は、本発明の実施形態に従ったメッセージ交換のシーケンスを示す図である。

ホストがドライバ（DRIVER__A）を起動したとき、初期化の一部として、SEND／RECEIVE 機構を利用可能とし、RDMA 機構を利用不可能とする。SEND／RECEIVE 機構は、NIC ドライバがリモートノードと通信できるように利用可能とされる。RDMA 機構は、あらゆる入力する（古い）RDMA 動作を拒絶するために、利用不可能とされる。

【 0 0 2 4 】

DRIVER__A は、システム全体に、起動していることを示すメッセージ（BOOTING）をブロードキャスト（あるいは、ユニキャスト）する。他の NIC ドライバから、その存在を認識したというメッセージを得るまで待つ。各ホストの応答が一回のみカウントされるように、応答を監視する。

【 0 0 2 5 】

DRIVER__A からの BOOTING メッセージを受信した NIC ドライバ（DRIVER__B）は、ローカルに保持している翻訳保護テーブルの DRIVER__A の部分をクリアする。テーブルの DRIVER__A の部分をクリアすることは、DRIVER__A の NIC に対する RDMA アクセスを禁止する効果がある。また、DRIVER__A の翻訳保護テーブルへの更新を送信することも禁止する。（DRIVER__B が、より早くに、DRIVER__A の NIC が機能しておらず、既にそれへのアクセスが禁止されていることを認識することは可能である。アクセスが既に禁止されているか否かは、ここでは問題ではない。）次のステップでは、DRIVER__B が、その NIC への全ての発行された動作、SEND あるいは RDMA が実行されたか否かを確認する。その後、DRIVER__B は、アクノレジメントメッセージ（ACK）で応答する。

【 0 0 2 6 】

DRIVER__A は、DRIVER__B に BOOTING メッセージを送信し、DRIVER__B が ACK メッセージで応答したので、DRIVER__A と D

R I V E R __ B との間に、まだ終わっていない、あるいは、実行中の R D M A 動作が無いことを保証する。(追い越しなし、及び、同じ経路を使っていることを仮定。) 換言すれば、D R I V E R __ A の N I C は、古い R D M A 動作から保護されている。D R I V E R __ B が、A C K を送る前に、全ての発行された動作が実行されたことを確認しないならば、この保証は成り立たない。A C K が、以前に発行された R D M A 動作の前に送られることはありえる。特に、N I C が発行された動作を処理するのに、複数のキューを持っている場合、これがあり得る。

【 0 0 2 7 】

リモート N I C が接続されていない、あるいは、動作していない場合、ドライバの A C K ではなく、ハードウェアで生成された (H W) N A C K が受信される。従って、ドライバは、システム内の全てのノードを数えることができ、どの N I C がシステムの一部で、どれがそうでないかを知ることができる。全ての N I C が数えられた時にのみ、D R I V E R __ A は、R D M A 機能を利用可能にする。そして、R D M A アクセスの準備ができたことを示すメッセージ (R D M A __ R E A D Y) を分かっている全ての N I C に送信する。

【 0 0 2 8 】

R D M A __ R E A D Y メッセージを受信する N I C ドライバ (D R I V E R __ B) は、その N I C への R D M A アクセスを利用可能にする。言い換えれば、D R I V E R __ A のシステム保護テーブルに、翻訳保護テーブルの更新を送信し始める。そして、D R I V E R __ A の保護テーブルの内、D R I V E R __ B の N I C を参照する部分を更新する。全てのリモート N I C ドライバが、D R I V E R __ A のテーブルのこれらの部分を更新したとき、D R I V E R __ A は、システムに完全に再組み込みされる。

【 0 0 2 9 】

ホストの状態は、何時変わるかわからないので、ドライバがそのホストのクラッシュに応答する前に、N I C ドライバ (D R I V E R __ C) によって B O O T I N G メッセージが受信されることもあり得る。この場合、D R I V E R __ A は、D R I V E R __ C の N I C を数えることができない。従って、D R I V E R __ A は、適切なタイムアウト時間後、数えることができなかった N I C に B O O T

INGメッセージを再送する。この処理は、全てのNICが数えられるまで繰り返される。別の方法として、DRIVER_CからACKメッセージの代わりに、BOOTINGメッセージを受け取った場合には、DRIVER_Aは、このプロセスをやらないことも可能である。

【0030】

新しいホストは、再起動したホストと同様の方法でシステムに加えることが可能である。

図5は、本発明の実施形態に従った、TPTにユーザプロセスが使う領域をユーザプロセスが登録する処理のフローチャートである。

【0031】

ステップS1においては、ノードNのユーザプロセスは、ドライバに登録されるべき領域の情報を与える。ステップS2においては、ノードNのドライバは、領域を特定し（あるいは、割当て）、論理アドレスを物理アドレスに変換し、ローカルTPTにエントリiを割当て、その領域に関する情報を格納する。ステップS3においては、ノードNのドライバは、ネットワークにつながれた全てのノードに更新要求を発行する。更新要求により、ノードNのドライバによって割り当てられた領域に対応する各ノードのTPTのエントリiを更新する。

【0032】

図6は、本発明の実施形態に従った、ユーザプロセスがRDMA Write 転送コマンドを発行する処理のフローチャートである。

ステップS10において、ユーザプロセスは、ローカルに登録された領域X（以下、RR_X）から、遠隔に登録された領域Y（以下、RR_Y）へBバイトデータ（Bは、バイト単位のデータ長である）を転送する要求である、RDMA コマンドをドライバへ発行する。ステップS11においては、ドライバは、ユーザプロセスが、ReadアクセスをRR_Xに発行したか否かを判断する。ステップS11の判断がNOの場合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップS11の判断がYESの場合には、ステップS12に進む。ステップS12においては、ドライバは、ユーザプロセスがRR_Yに対し、Writeアクセスを発行したか否かを判断する。ステップS12の判断がNOの場合

合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップ S 1 2 の判断が Y E S の場合には、ステップ S 1 3 に進む。ステップ S 1 3 においては、ドライバは、転送されたデータの長さが、2 つの領域（すなわち、R R _ X と R R _ Y）の限界以内であるか否かを判断する。ステップ S 1 3 の判断が N O の場合には、ドライバは、ユーザプロセスに保護エラーを返す。ステップ S 1 3 の判断が Y E S の場合には、ドライバは、N I C にデータを転送させるため、N I C に対し、転送コマンドを発行する。

【 0 0 3 3 】

図 7 は、本発明の実施形態に従った、ホストの起動処理を示すフローチャートである。

ホストが起動された後、ドライバは、N I C と T P T を初期化し、S E N D / R E C E I V E 機能を利用可能とし、R D M A 機能を利用不可能とし、{応答集合} という変数を全ての 0 に設定する（ステップ S 1 5）。ステップ S 1 6 においては、ドライバは、全てのノードに B O O T I N G メッセージを送信する。ステップ S 1 7 では、ドライバは、所定の時間内で、他のノードからの応答を待つ。ステップ S 1 7 でタイムアウトが起こったときは、ステップ S 1 6 に戻る。ステップ S 1 7 において、ドライバが、応答メッセージ（A C K、ハードウェアで生成された N A C K、B O O T I N G メッセージのいずれか）を受信した場合には、取り阿波は、応答メッセージの発信ノードを特定し、ステップ S 1 8 において、{応答集合} にノードの識別子を加える。

【 0 0 3 4 】

例えば、応答集合は、0 と 1 の列から構成される。ステップ S 1 5 においては、この列は全て 0 に設定される。この列中、各位は、ホストに対応付けられる。自ホストのドライバによって応答が受信された場合には、応答を送信したホストが特定され、対応する位置にある列内の 0 を 1 に設定する。列が 1 で満たされた場合、これは、全てのホストから応答を受信したことを示す。

【 0 0 3 5 】

ステップ S 1 9 において、ドライバは、全てのノードが {応答集合} にあるか否かを判断する。ステップ S 1 9 における判断が、N O の場合には、ステップ S

17に進む。ステップS19の判断がYESの場合には、ステップS20に進む。ステップS20では、ドライバは、RDMA機能を利用可能とし、ステップS21において、全てのノードにRDMA_READYメッセージを送信する。

【0036】

図8は、起動されたノード以外のノードの動作を示すフローチャートである。

ステップS25において、ドライバは、起動されたノードNのドライバからBOOTINGメッセージを受信する。ステップS26において、ドライバはノードNのTPTエントリをクリアし、ノードNへのTPT更新を利用不可能にする。ステップS27においては、ドライバは、ノードNに対し、ACKで応答する。他の場合には、ノードN以外のノードのドライバは、ダウンしている可能性がある。この場合、BOOTINGメッセージへの応答(HW NACK)は、ハードウェアで生成され、ノードNに送り返される。また、ノードN以外のノードは、ちょうど起動されたばかりかもしれない。この場合、BOOTINGメッセージがノードNに送られ、このノードがノードNとして動作する。

【0037】

いずれにしても、ドライバは、ステップS28において、RDMA_READYメッセージをノードNから受信する。ステップS29においては、ドライバは、ノードNに対するTPT更新を利用可能とする。そして、ステップS30において、ドライバは、そのローカルTPT部の内容により、ノードNのTPTを更新する。この更新は、例えば、RDMA Writeコマンドにより達成される。

【0038】

処理フローは、2つのホストのみを参照して説明したが、この処理フローは、多くのホスト間で実行される。

(付記1) RDMA機能を有した、ホストと複数のホストからなるネットワーク内の該ホストに設けられた装置であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第1のメッセージを、該ネットワーク内の全ての該複数のホストに送信する手段と

該複数のホストからの該ホストへの R D M A アクセスを利用不可能とする手段と、

該ホストに第 2 のメッセージを送信することによって、第 1 のメッセージに回答する手段と、

該ホストが、該複数のホストからの R D M A アクセスを受付可能であることを示す第 3 のメッセージを、該複数のホストの全てから第 2 のメッセージを受信し、R D M A 機能を利用可能とした後、該複数のホストの全てに送信する手段と、を備えることを特徴とする装置。

【 0 0 3 9 】

(付記 2) 前記装置は、前記ホストのドライバに含まれることを特徴とする付記 1 に記載の装置。

(付記 3) 他のホストに R D M A アクセスを行うための情報を有する翻訳保護テーブル手段を更に有し、

前記第 1 のメッセージを受信したとき、該第 1 のメッセージを送信したホストに関する情報を該翻訳保護テーブル手段からクリアし、該ホストへの R D M A アクセスを不可能にすることを特徴とする付記 1 に記載の装置。

【 0 0 4 0 】

(付記 4) 前記複数のホストの翻訳保護テーブルは、該複数のホストへ、前記第 3 のメッセージが送られた後、更新されることを特徴とする付記 3 に記載の装置。

【 0 0 4 1 】

(付記 5) 前記第 2 のメッセージは、アクノレジメント、ノンアクノレジメント、及び、前記複数のホストから送られた前記第 1 のメッセージの一つであり、ノンアクノレジメントは、ハードウェアによって生成されることを特徴とする付記 1 に記載の装置。

【 0 0 4 2 】

(付記 6) 前記第 2 のメッセージが前記複数のホストの全てから受信されたか否かは、監視され、0 と 1 の列からなる応答集合によって判断されることを特徴とする付記 1 に記載の装置。

【 0 0 4 3 】

(付記 7) 前記ホストは、R DMA 機能と他のメッセージ通信機能を有するネットワークインターフェースカードを有し、R DMA 機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする付記 1 に記載の装置。

【 0 0 4 4 】

(付記 8) R DMA 機能を有した、ホストと複数のホストからなるネットワーク内の該ホストの方法であって、

該ネットワーク内のホストが起動したとき、該ホストが起動したことを示す第 1 のメッセージを、該ネットワーク内の全ての該複数のホストに送信するステップと、

該複数のホストからの該ホストへの R DMA アクセスを利用不可能とするステップと、

該ホストに第 2 のメッセージを送信することによって、第 1 のメッセージに回答するステップと、

該ホストが、該複数のホストからの R DMA アクセスを受付可能であることを示す第 3 のメッセージを、該複数のホストの全てから第 2 のメッセージを受信し、R DMA 機能を利用可能とした後、該複数のホストの全てに送信するステップと、

を備えることを特徴とする方法。

【 0 0 4 5 】

(付記 9) 前記方法は、前記ホストのドライバで行われることを特徴とする付記 8 に記載の方法。

(付記 1 0) 他のホストに R DMA アクセスを行うための情報を格納する翻訳保護テーブルステップを更に有し、

前記第 1 のメッセージを受信したとき、該第 1 のメッセージを送信したホストに関する情報を該翻訳保護テーブルステップで格納された情報からクリアし、該ホストへの R DMA アクセスを不可能にすることを特徴とする付記 8 に記載の方法。

【 0 0 4 6 】

(付記 1 1) 前記複数のホストの翻訳保護テーブルは、該複数のホストへ、前記第 3 のメッセージが送られた後、更新されることを特徴とする付記 1 0 に記載の方法。

【 0 0 4 7 】

(付記 1 2) 前記第 2 のメッセージは、アクノレジメント、ノンアクノレジメント、及び、前記複数のホストから送られた前記第 1 のメッセージの一つであり、ノンアクノレジメントは、ハードウェアによって生成されることを特徴とする付記 8 に記載の方法。

【 0 0 4 8 】

(付記 1 3) 前記第 2 のメッセージが前記複数のホストの全てから受信されたか否かは、監視され、0 と 1 の列からなる応答集合によって判断されることを特徴とする付記 8 に記載の方法。

【 0 0 4 9 】

(付記 1 4) 前記ホストは、R D M A 機能と他のメッセージ通信機能を有するネットワークインターフェースカードを有し、R D M A 機能と他のメッセージ通信機能の初期化は独立して行われることを特徴とする付記 8 に記載の方法。

【 0 0 5 0 】

【発明の効果】

本発明によれば、R D M A 機能を有する N I C を持つホストを、ユーザレベルプロセスを用いずに、ネットワークに安全に組み込む装置が提供される。

【図面の簡単な説明】

【図 1】

本発明の実施形態が基本とするシステム構成を示す図である。

【図 2】

ホストの構成を示す図である。

【図 3】

翻訳保護テーブル (translation and protection table) の構成を示す図である。

【図 4】

本発明の実施形態に従った、メッセージ交換のシーケンスを示す図である。

【図 5】

本発明の実施形態に従った、ユーザプロセスが、ユーザプロセスが使用する領域を T P T に登録する処理のフローチャートである。

【図 6】

本発明の実施形態に従った、ユーザプロセスが R D M A 転送コマンドを発行する処理のフローチャートである。

【図 7】

本発明の実施形態に従った、ホストの起動処理のフローチャートである。

【図 8】

起動されたノード以外のノードの動作を示すフローチャートである。

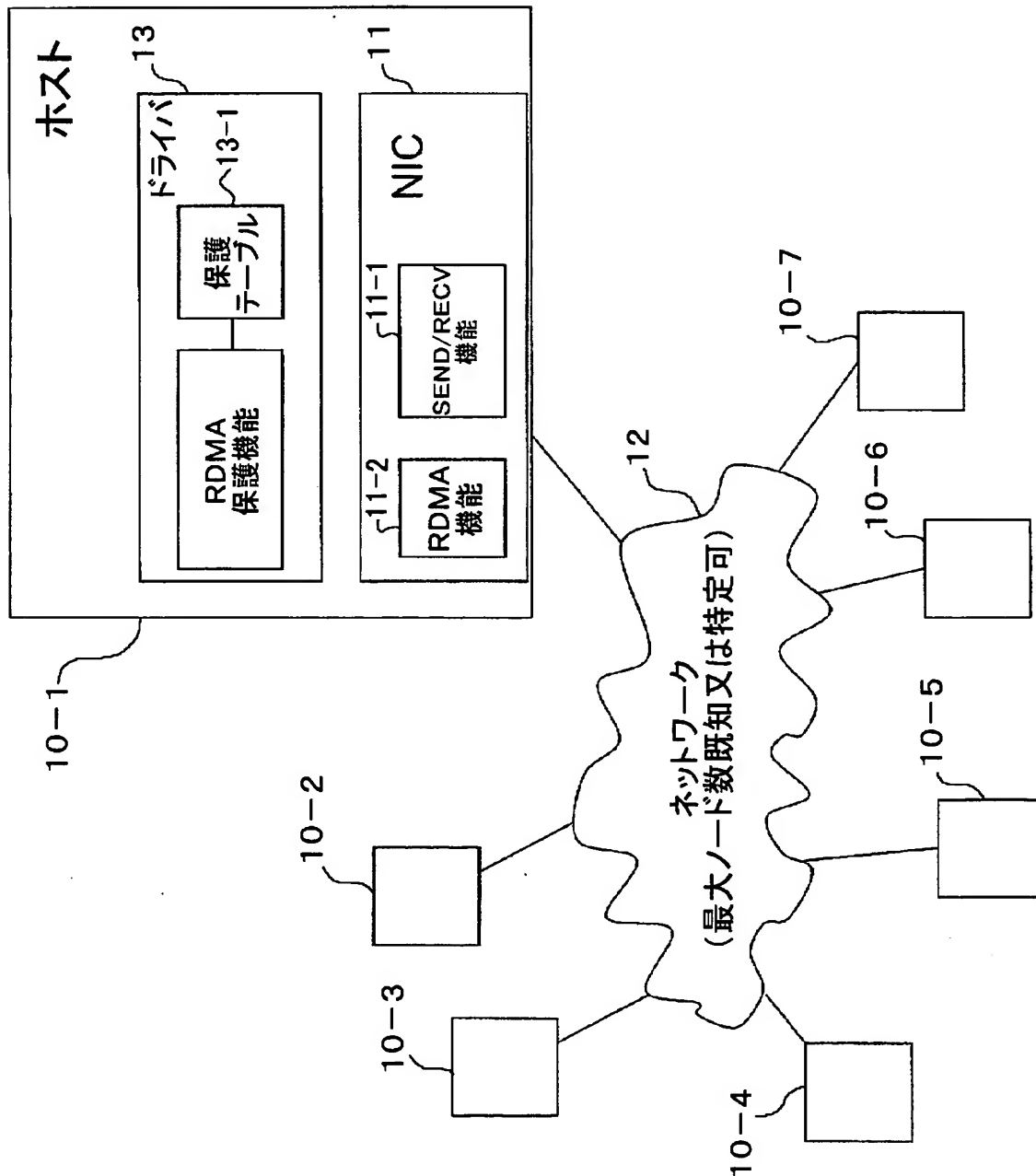
【符号の説明】

1 0 - 1 ~ 1 0 - 7、1 0 - i、1 0 - j	ホスト
1 1	N I C
1 1 - 1	S E N D / R E C E I V E 機能
1 1 - 2	R D M A 機能
1 2	ネットワーク
1 3	ドライバ
1 3 - 1	R D M A 保護機能
2 0	ユーザ空間
2 1	カーネル空間

【書類名】 図面

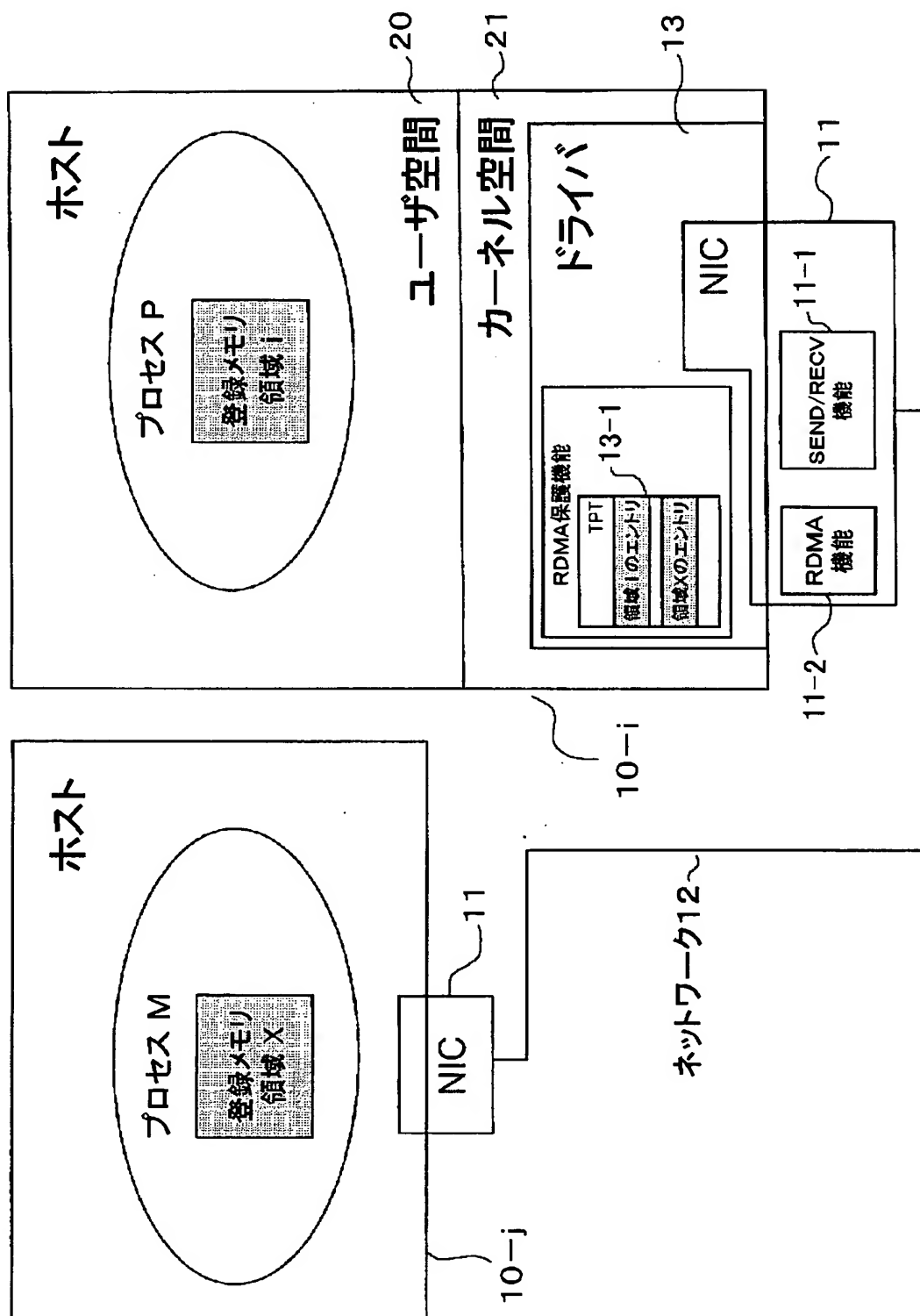
【図 1】

本発明の実施形態が基本とするシステム構成を示す図



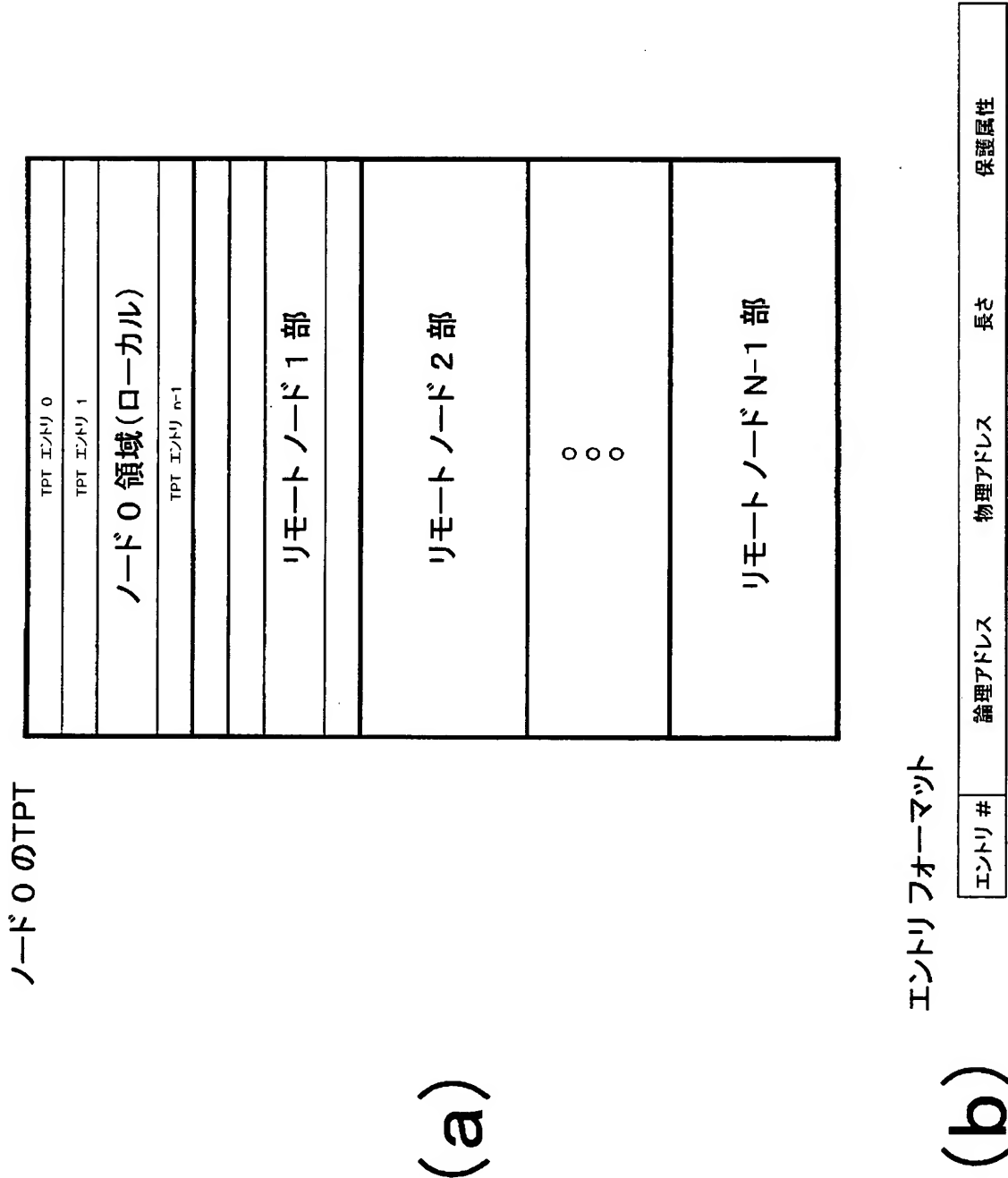
【図 2】

ホストの構成を示す図



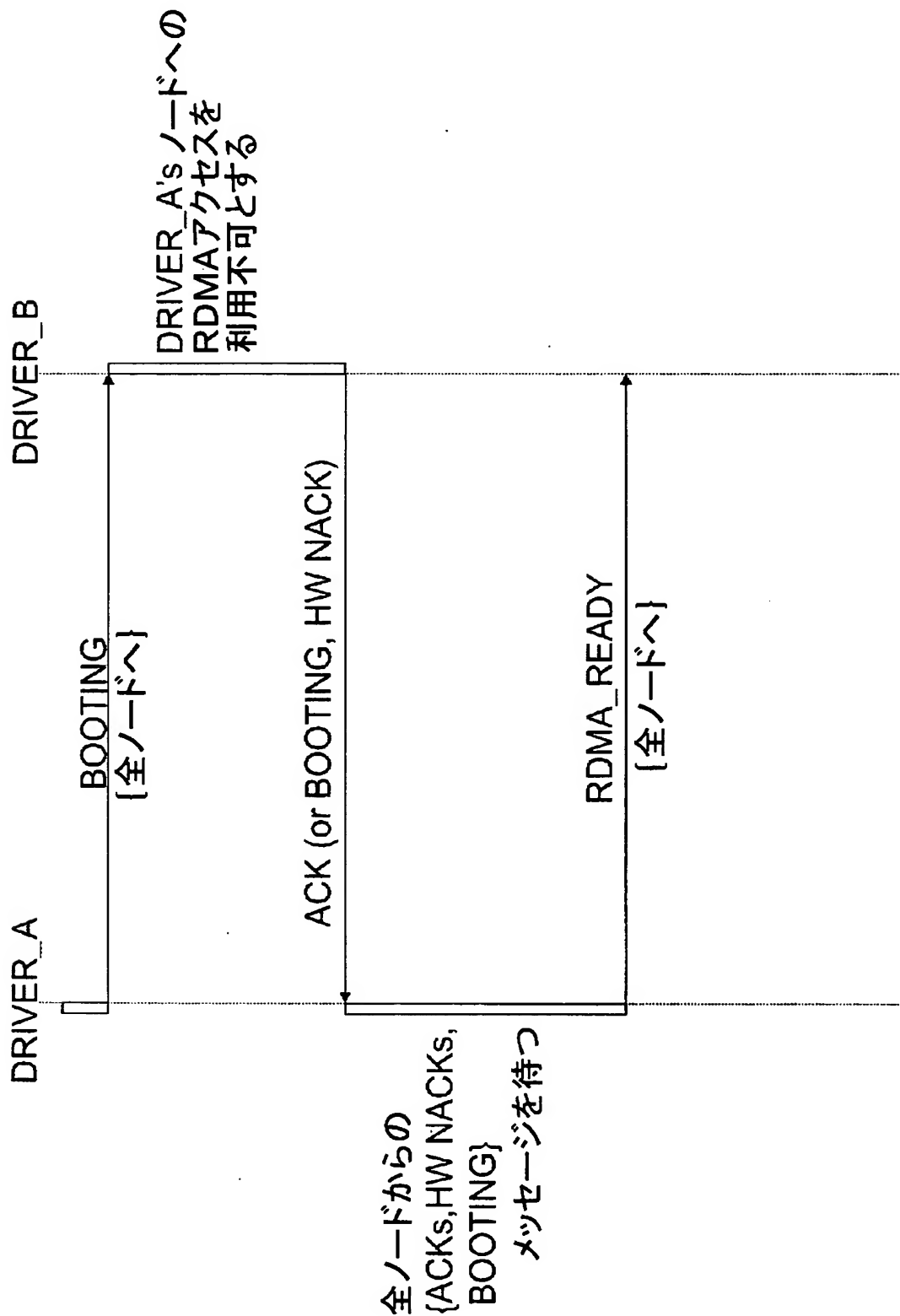
【図 3】

翻訳保護テーブル (translation and protection table)
の構成を示す図



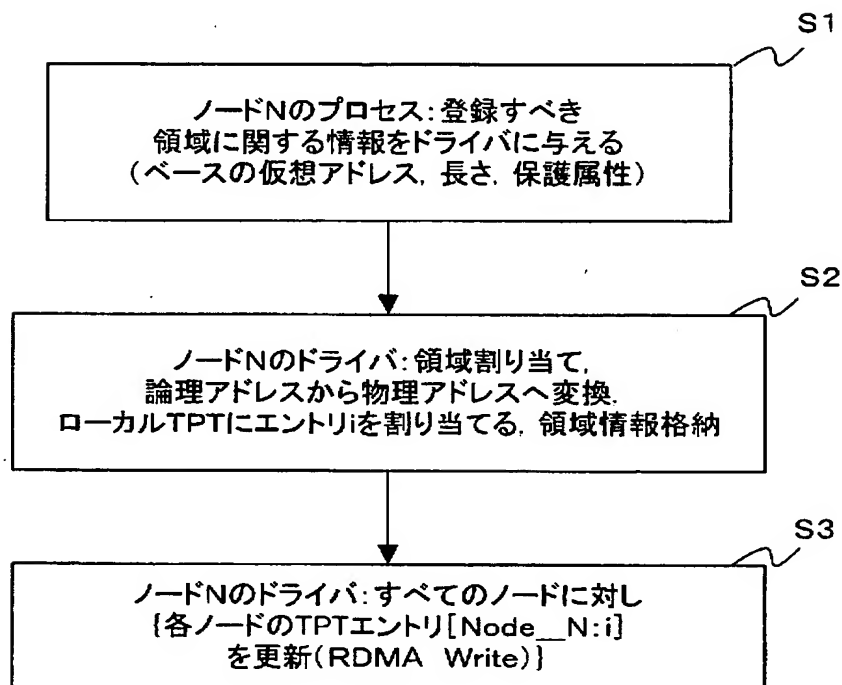
【図 4】

本発明の実施形態に従った、
メッセージ交換のシーケンスを示す図



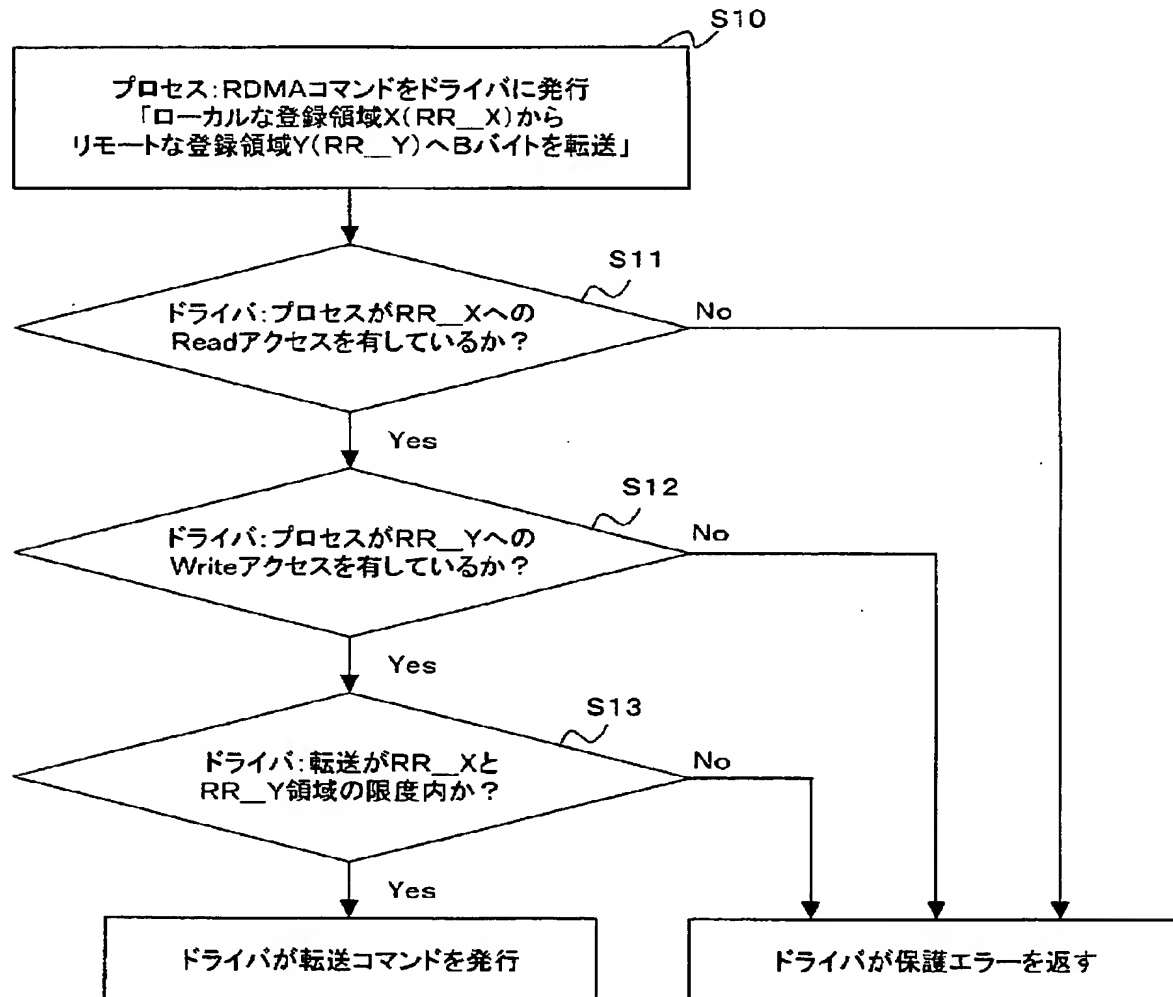
【図 5】

本発明の実施形態に従った、ユーザプロセスが、
ユーザプロセスが使用する領域を
TPTに登録する処理のフローチャート

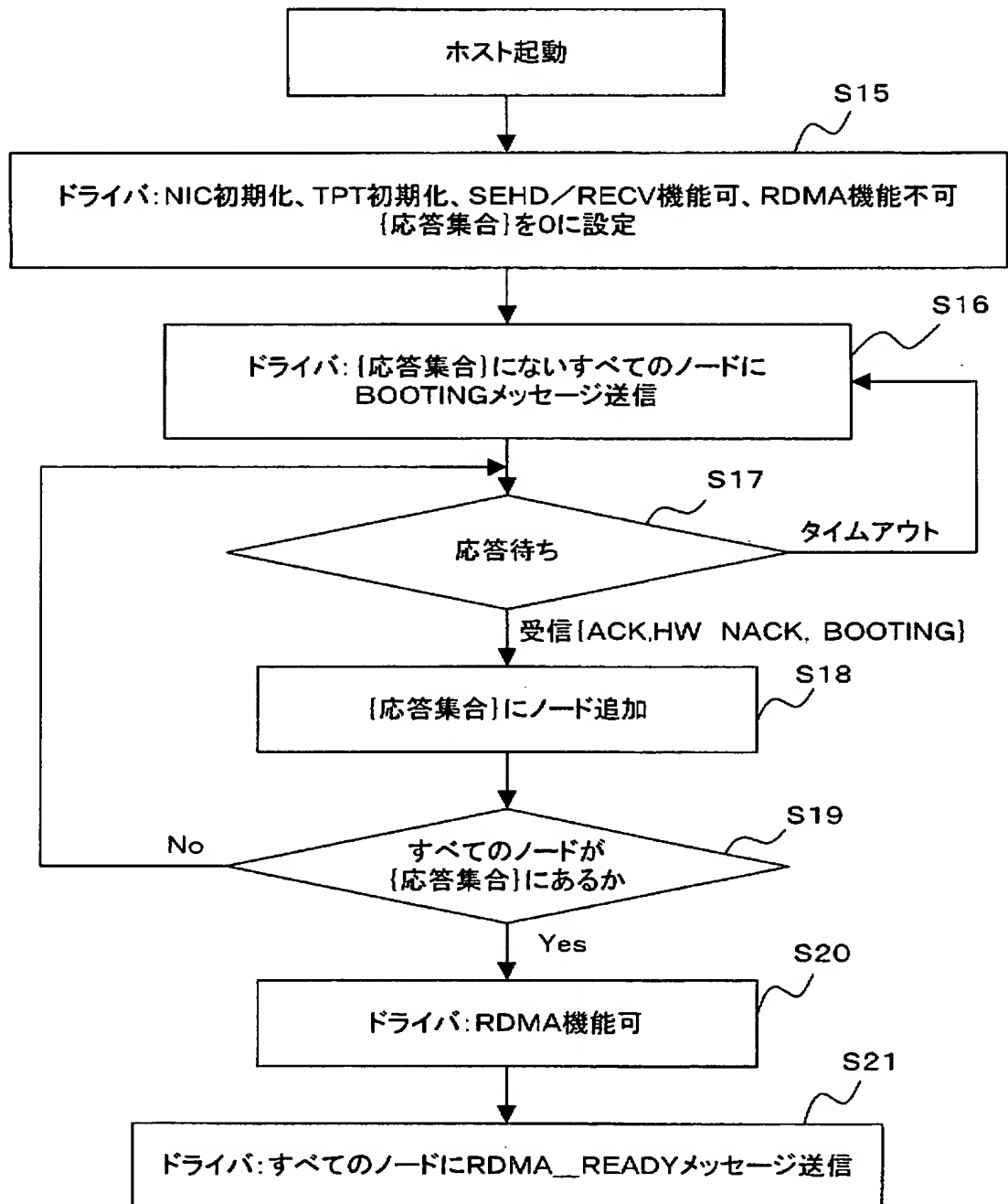


【図 6】

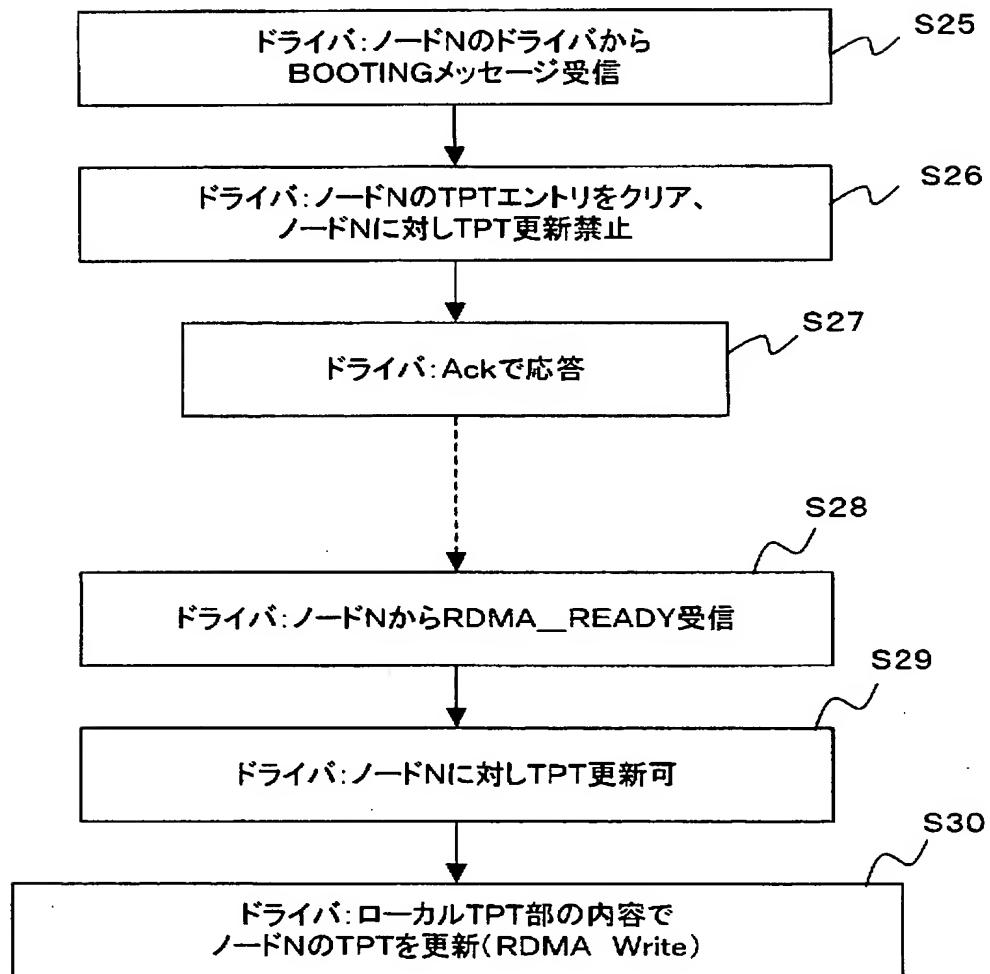
本発明の実施形態に従った、ユーザプロセスが
RDMA転送コマンドを発行する処理のフローチャート



【図 7】

本発明の実施形態に従った、
ホストの起動処理のフローチャート

【図 8】

起動されたノード以外の
ノードの動作を示すフローチャート

【書類名】 要約書

【要約】

【課題】 R D M A 機能を有する N I C を持つホストをネットワークに安全に組み込むための装置を提供する。

【解決手段】 複数のホストを含むネットワークにおいて、ホストが起動すると、該ホストのドライバは、該ホストが起動することを示す B O O T I N G メッセージを全ての他のホストに送り、これを受信した、起動するホスト以外のホストのドライバは、起動するホストへの R D M A アクセスを利用不可能にする。起動するホスト以外のホストのドライバは、A C K あるいは B O O T I N G メッセージで応答する。ある場合には、ネットワークがハードウェア N A C K を生成する。起動するホストが全てのホストから応答を受信すると、起動するホストのドライバは自ホストへの R D M A アクセスを許可する。

【選択図】 図 4

特願 2 0 0 2 - 3 5 7 4 4 9

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 2 2 3]

1. 変更年月日
[変更理由]

1 9 9 6 年 3 月 2 6 日

住所変更

住 所
氏 名

神奈川県川崎市中原区上小田中 4 丁目 1 番 1 号
富士通株式会社